ENCOURAGING DEMAND PARTICIPATION IN TEXAS' POWER MARKETS

prepared by

Dr. Steven Braithwait
Dr. B. Kelly Eakin
Dr. Laurence D. Kirsch
Laurits R. Christensen Associates, Inc.

for

The Market Oversight Division of the Public Utility Commission of Texas

August 31, 2002

TABLE OF CONTENTS

EXECUTIVE SUMMARY .... **III**

1. INTRODUCTION **1**
- 1.1. Purpose and Scope of Project and Technical Report 1
- 1.2. Process of Getting Stakeholder Input 1
- 1.3. Organization of the Report 2

2. GENERAL BENEFITS OF INCORPORATING DEMAND RESPONSE INTO WHOLESALE MARKETS **2**
- 2.1. Achieving Efficient Demand Response 3
  - 2.1.1. Communicating Market Conditions to Consumers 3
  - 2.1.2. Knowledgeable Consumer Response 6
  - 2.1.3. Measuring Consumers' Responses 6
- 2.2. Market-Based Pricing and Customer Choice 6

3. TYPES OF DEMAND RESPONSE AND MARKETS FOR DEMAND RESPONSE **7**
- 3.1. Types of Demand Response 7
- 3.2. Markets for Demand Response 8
  - 3.2.1. Energy Markets 9
  - 3.2.2. Ancillary Services Markets 9
- 3.3. Barriers to Demand Response 10
  - 3.3.1. Customer Barriers 10
  - 3.3.2. Technological Barriers 10
  - 3.3.3. Regulatory Barriers 10
  - 3.3.4. Load Profiling 11
  - 3.3.5. Quantifying the Amount of Load Response 12

4. REVIEW OF DEMAND RESPONSE PROGRAMS IN OTHER ISOS **12**
- 4.1. PJM Interconnection 13
- 4.2. New York ISO 13
- 4.3. ISO New England 13
- 4.4. California ISO 13
- 4.5. Other Demand Response Initiatives 14

5. DEMAND PARTICIPATION IN ERCOT'S MARKETS **14**
- 5.1. Load Participation Opportunities in ERCOT 15
  - 5.1.1. Passive Load Response 15
  - 5.1.2. Certified Balancing Up Load Resource 16
  - 5.1.3. Load as a Resource 16
- 5.2. Current Extent of Load Participation in ERCOT Markets 16
- 5.3. Barriers to Demand Participation in ERCOT Markets 17
  - 5.3.1. Major Barriers to Demand Participation 17
  - 5.3.2. Impediments Limiting Demand Participation 18
  - 5.3.3. Potential Barriers 19

6. EFFECTS OF EXISTING ERCOT MARKET RULES ON EFFICIENT DEMAND RESPONSE 19
6.1. Signaling Consumers About Market Conditions 20
6.1.1. Energy Markets 20
6.1.2. Ancillary Service Markets 20
6.2. Consumers' Responses to Market Signals 21
6.3. Measuring the Services Provided by Consumers 22

7. RECOMMENDATIONS 23
7.1. Short-Term Recommendations 24
7.1.1. Select a Standard Method for Measuring the "Baseline Loads" of Curtailed Customers 24
7.1.2. Survey Customers, REPs and QSEs on BUL and LaaR on Their Demand Response Program Participation Decisions 24
7.1.3. Develop Pilot Curtailment Programs 25
7.1.4. Develop Benchmarks of the Benefits of Demand Response Programs 26
7.1.5. Evaluate Metering Policy 27
7.1.6. Improve the BUL and LaaRs Document 27
7.1.7. Require Each Resource to Pay Its Own Overhead Costs 27
7.1.8. Evaluate the Benefits and Costs of Our Long-Term Recommendations 27
7.2. Intermediate-Term Recommendations 28
7.2.1. Develop Non-Discriminatory Measures of Performance 28
7.3. Long-Term Recommendations 30
7.3.1. Develop Transparent Day-Ahead Electricity Markets 31
7.3.2. Develop Transparent Same-Day Electricity Markets 31
7.3.3. Consider Adopting Three-Part Bidding 32
7.3.4. Consider Adopting Efficient Locational Pricing of Electrical Energy 32
7.3.5. Apply Market-Clearing Prices to All Resources' Effective Quantities 33
7.3.6. Set Penalties for Non-Performance Equal to the Costs of Non-Performance 33

ATTACHMENT 1. ERCOT DOCUMENTS REVIEWED 35

EXECUTIVE SUMMARY

The Market Oversight Division (MOD) of the Public Utility Commission of Texas (PUC) has requested an analysis and critical evaluation of the interactions between the ERCOT market rules and demand participation. We have undertaken these tasks with the belief that the goal of demand response programs should be to maximize the net benefits that consumers derive from electricity. We believe that a net benefit test is important for differentiating programs that create social wealth from programs that destroy social wealth.

Demand response mechanisms may be divided into three generic categories: dynamic pricing (known in Texas as passive load response), load reduction programs (like interruptible service), and ancillary service programs. Although each of these categories of demand response mechanisms can help connect the wholesale and retail energy markets, they differ in whether and how they measure demand response, communicate demand response to the wholesale market, pay consumers for demand response.

The barriers to demand resource participation are of three sorts. *Customer barriers* include risk aversion and inflexibility in consumption. *Technological barriers* include limitations in metering, communication, and software. *Regulatory barriers* include consumer protection schemes that prevent consumers from receiving wholesale market signals or from having incentives to respond to those signals. *Measurement barriers* arise from the inadequacy of present methods for determining the quantity of each customer's demand response, even with adequate metering.

Each of the four existing FERC-regulated ISOs – the Pennsylvania-Maryland-New Jersey Interconnection (PJM), New York, New England, and California – have demand response programs in place. These programs fall into two general categories: "emergency programs" that respond to system reliability concerns; and "economic programs" that capture economic benefits according to wholesale market price levels. Unfortunately, the performance of these programs in the summer of 2001 was modest at best, and more often insignificant. Their limited performance likely resulted from the newness of the programs, their design features, and barriers to customer acceptance.

We have found that demand resources have limited opportunities to participate in wholesale markets in ERCOT. The actual opportunities for demand participation in ERCOT currently appear to be only in the ancillary services markets. This is problematic because energy markets provide the greatest potential source of benefits from demand participation but the opportunities to achieve these benefits are severely limited in ERCOT.

We have discovered several factors that appear to limit demand participation in wholesale markets in Texas. The *major barriers* to demand participation in ERCOT are:

- absence of centralized day-ahead and same-day energy markets;
- the balanced schedule requirement; and
- the degrees of separation between the customer loads and the wholesale markets.

The secondary group of barriers, *impediments limiting demand participation*, includes:

- restrictive and complicated certification process for BULs and LaaRs;
- difficulties in accurately measuring load response; and
- lack of interval metering and reliance on load profiling, which together create disincentives for offering demand response opportunities.

The tertiary group, *potential barriers*, identifies factors that might limit future demand responsiveness, but that currently are not limiting because of the effect of the barriers in the first two groups. These potential barriers are:

- non-market based standard service offers (e.g., the "price-to-beat"); and
- inefficient transmission and distribution pricing.

We have reviewed the effects of ERCOT market rules on demand participation and make the following three sets of recommendations:

Short-Term Recommendations

S1. Select a standard method for measuring the "baseline loads" of curtailed customers
S2. Survey customers, REPs and QSEs on BUL and LaaR on their demand response program participation decisions
S3. Develop pilot curtailment programs
S4. Develop benchmarks of the benefits of demand response programs
S5. Evaluate metering policy
S6. Improve the BUL and LaaRs document
S7. Require each resource to pay its own overhead costs
S8. Evaluate the benefits and costs of our long-term recommendations

Intermediate-Term Recommendations

I1. Develop non-discriminatory measures of performance

Long-Term Recommendations

L1. Develop transparent day-ahead electricity markets
L2. Develop transparent same-day electricity markets
L3. Consider adopting three-part bidding
L4. Consider adopting efficient locational pricing of electrical energy
L5. Apply market-clearing prices to all resources' effective quantities
L6. Set penalties for non-performance equal to the costs of non-performance

Each of these recommendations, and their underlying rationales, is explained in the text.

ENCOURAGING DEMAND PARTICIPATION IN TEXAS' POWER MARKETS

Dr. Steven Braithwait
Dr. B. Kelly Eakin
Dr. Laurence D. Kirsch
Laurits R. Christensen Associates, Inc.

1. INTRODUCTION

The Electric Reliability Council of Texas (ERCOT) is in the midst of restructuring its electricity market. Because many parties to this market recognize that demand participation can play a crucial role in insuring market stability, mitigating market power, and maximizing the net benefits that electricity brings to Texans, the Market Oversight Division (MOD) of the Public Utility Commission of Texas (PUC) has requested an analysis and critical evaluation of the interactions between the ERCOT market rules and demand participation.

1.1. Purpose and Scope of Project and Technical Report

The study has three main objectives. The first is to review how the existing ERCOT market rules and processes affect demand participation and price responsiveness in energy and ancillary services markets. The second objective is to identify possible modifications to those market rules to encourage efficient demand participation and price responsiveness in ERCOT's markets. The third objective is to identify programs for encouraging demand participation and price responsiveness in ERCOT markets. These objectives reflect the fact that price-responsive retail load is imperative for the efficient operation of wholesale power markets. These objectives are also responsive to the mandate of the PUC for ERCOT to "develop measures and refine existing measures, to enable load resources a greater opportunity to participate in the ERCOT markets." [1]

1.2. Process of Getting Stakeholder Input

Obtaining stakeholder input has been a central component of this project. The project was initiated with a conference call with MOD staff. In this call, the general approach and deliverables (as delineated in the Agreement for Services) for the project were reviewed and agreed upon. Based on this call, a presentation was drafted and presented (via teleconference) to the June 3, 2002, meeting of the ERCOT Demand Side Working Group (DSWG). Feedback from the stakeholders was obtained from the DSWG attendees. In addition, numerous documents were provided to the project team by the MOD and the DSWG.[2]

Based upon all of the foregoing information and feedback, an interim memorandum entitled "Encouraging Demand Participation in Texas" Power Markets" was prepared and distributed to the DSWG. The memorandum and an accompanying presentation were discussed at the

---

[1] Docket No. 23220, *Petition of the Electric Reliability Council of Texas for Approval of the ERCOT Protocols*, June 4, 2001, pp. 20-21.

[2] A list of the documents which have been received and reviewed appears as Attachment 1.

July 8, 2002, meeting of the DSWG. Following that meeting, the project team received written comments from the MOD staff and several members of the DSWG. In addition, there have been follow-up conference calls with MOD staff. This technical report has attempted to incorporate the insights gained from the attendance at the July 8 DSWG meeting as well as to address the issues received from stakeholder and MOD staff feedback.

1.3. Organization of the Report

The report is organized as follows. Section 2 discusses the general benefits of incorporating retail demand response into wholesale markets. Section 3 describes different categories of demand response, the markets in which demand response may participate, and certain key barriers that tend to inhibit demand participation in wholesale markets. Section 4 presents a brief review of demand response programs in other ISOs. Section 5 addresses the issues of how demand participation occurs in ERCOT markets and what are the key barriers to greater demand participation. Section 6 assesses the effect of the existing ERCOT market rules on demand participation. Finally, Section 7 recommends programs and studies that Texas can undertake in the short- to intermediate-term time frames, plus certain changes in the ERCOT market rules that can be implemented in a longer time frame.

2. GENERAL BENEFITS OF INCORPORATING DEMAND RESPONSE INTO WHOLESALE MARKETS

The goal of demand response programs should be to maximize the net benefits that consumers derive from electricity, including environmental benefits. This means that customers should generally be encouraged to *reduce* load when the incremental value of their load is *less* than the incremental cost of serving their load, and that they should generally be encouraged to *increase* load when the incremental value of their load is *more* than the incremental cost of serving their load; and it also means that these demand responses should be encouraged only when the costs of administering and metering these responses is less than the social benefits of these responses. [3]

This net benefit test is important for differentiating desirable demand response programs from undesirable programs. A demand response program that has benefits greater than costs creates wealth. A demand response program that has benefits less than costs destroys wealth. The first kind of program can pay for itself. The second kind of program can only exist if it is subsidized by taxpayers or by consumers who pay more for the program than they receive.

Demand response to prices that reflect wholesale market conditions has several important benefits. In particular, it can relieve transmission constraints, reduce the severity of wholesale price spikes, reduce the potential exercise of market power by generators, and offer cheap power to consumers when supplies are especially plentiful. Although the benefits in individual high-priced hours can be very large, these hours typically represent a very small fraction of a year. While the efficiency gain in a typical low-priced hour may be small, there are vastly more hours

[3] The goal of maximizing net benefits is not the same as maximizing the number of customers who participate in demand-side programs, or maximizing demand-side program participation rates, or maximizing the MWs of demand response. Demand-side management (DSM) programs have a history of flaws and pitfalls involving inefficiency, free-ridership, and cross-subsidies. Making these same mistakes today not only is costly today, but can hinder the on-going development of competitive electricity markets.

in which this gain can occur; so it is possible that, as electricity markets mature, getting market signals right in the low-priced hours may produce the greatest efficiency gains.[4]

2.1. Achieving Efficient Demand Response

Economically efficient demand response occurs when customers match the marginal value of their electricity consumption with the power system's marginal opportunity cost of electricity production.[5] Three broad steps are required to achieve such an efficient response:

1. Market conditions must be communicated to consumers through either a price signal or a quantity (*e.g.*, curtailment) signal. These signals should accurately reflect the marginal opportunity cost of electricity production, and should be communicated in a timely fashion.
2. Consumers' load responses must reflect a knowledgeable comparison of the price signal that they receive with the value of their consumption.
3. The system operator must accurately measure consumers' responses and arrange settlements accordingly.

Each of these steps is discussed below.

2.1.1. Communicating Market Conditions to Consumers

The marginal opportunity cost of electricity varies by electricity service, by time of day and year, and by power system location. Ideally, the power market would be designed so that wholesale market prices accurately reflected the marginal cost of each service, by time and location. With those prices serving as their incremental costs, retail electricity providers (REPs) would have a profit incentive to create efficient demand response programs.[6] If the resulting programs were attractive to consumers and profitable to the REP, they would also be beneficial to Texas as a whole.

If the market design yields inefficient wholesale prices that do not accurately reflect incremental costs, then creating efficient demand response programs is made more difficult. In such a case, developing an efficient demand response program requires that someone estimate the economic values of demand response by service, time, and location. If the inefficient market prices are less than these economic values, it will be beneficial to Texas to pay consumers to conserve electricity. This raises many logistical problems concerning the sources of money for these payments, the manner in which REPs will serve as conduits for these payments, and so on.

Accurate price signals must be communicated to customers in a timely manner. This must be done in a fashion that provides incentives for consumers to compare the benefits that they derive

---

[4] Consequently, rate forms such as two-part real-time pricing seem superior to "demand-side bidding" in that they benefit consumers in both surplus and supply-constrained situations.

[5] In principle, the marginal opportunity cost includes any externalities, such as pollution costs.

[6] REPs will arrange to buy much of their requirements in forward markets at fixed prices. Thus, their *financial costs* will reflect these contract prices. However, their *incremental opportunity costs* of balancing purchases or sales will reflect the wholesale market prices of the services that they trade at each time and location.

from electric services to the power system's costs of providing those services. Communication involves several key components:

1. determining efficient wholesale prices in a timely fashion;
2. giving appropriate incentives to REPs to communicate these prices to consumers through retail electricity products;
3. accurately measuring consumers' response to these prices; and
4. installing the hardware systems (like computers, metering, and communication equipment) that are required to facilitate calculation and communication.[7]

Under traditional regulation, the benefits of demand response were seldom realized because the market design did not allow wholesale price signals to reach consumers, nor did it allow consumers to express their willingness-to-pay for services in a manner that could be communicated to the wholesale market (e.g. by offering to reduce load in return for a financial payment tied to the wholesale price). This situation is illustrated in Figure 1, in which retail customers are represented as having demand that depends upon weather conditions but is nevertheless unresponsive to changing wholesale prices. The lack of retail demand response results from a fixed retail price. The impact in the wholesale market is that aggregate demand is effectively perfectly inelastic. When the demand increases (say due to hot weather) or supply decreases (say due to a generation unit outage), wholesale price spikes occur with no demand-side relief. The economic impacts are greatly increased costs and economic inefficiency as consumers use electricity that has a cost that greatly exceeds its benefit.

Introducing demand response into wholesale electricity markets leads to lower costs relative to consumer value, increased economic efficiency, and reduced price volatility. This is illustrated in Figure 2. As wholesale costs are conveyed to retail customers facing dynamic retail prices, the resulting load reduction allows the wholesale market to clear at lower prices. Furthermore, in tight capacity situations, the incremental cost of supply rises sharply with small increases in output; so a small amount of demand reduction can result in dramatic reductions in wholesale prices.

For example, Caves, Eakin, and Faruqui conducted a simulation showing that having about 10% of retail load on a real-time price would have mitigated the Midwest price spikes of 1998 and 1999 by about 60%.[8] Similarly, Braithwait and Faruqui estimated that if California had 50% of its large industrial load and 25% of its large commercial customer load on real-time pricing, a typical wholesale price spike in the range of $750/MWh would produce a load reduction of 2.5%, which would in turn cause a reduction in wholesale prices of 24%.[9] Along those same

---

[7] Such communication requirements apply equally to dynamic pricing programs and to load reduction (or "quantity programs") like curtailments.

[8] D. Caves, K. Eakin, and A. Faruqui, "Mitigating Price Spikes in Wholesale Markets through Market-Based Pricing in Retail Markets," *The Electricity Journal*, April 2000.

[9] S.D. Braithwait and A. Faruqui, "Demand Response – The Ignored Solution to California's Energy Crisis," *Public Utility Fortnightly*, March 15, 2001.

lines, Hirst and Kirby concluded that a 5% reduction in demand would have reduced the highest wholesale prices in California by 50%.[10]

Figure 1. Disconnected Electricity Markets

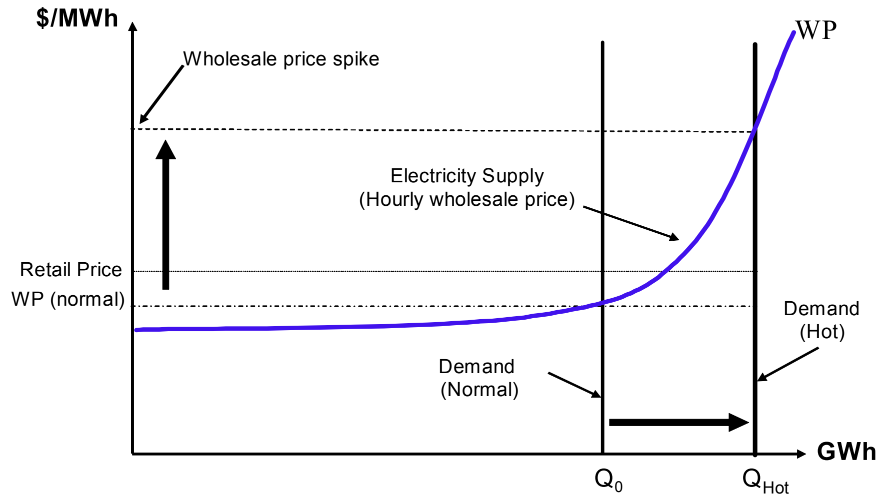
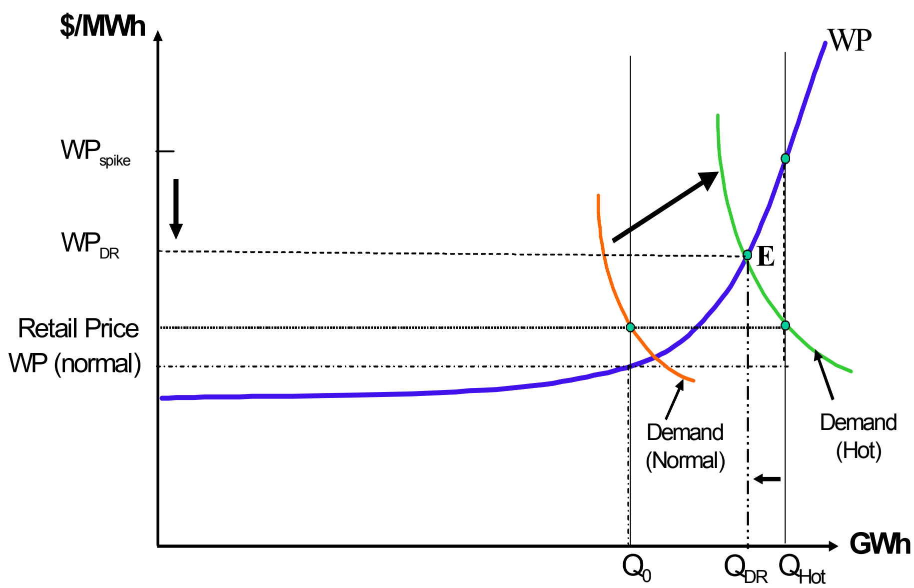


Figure 2. Effect of Demand Response on Wholesale Prices



[10] E. Hirst and B. Kirby, "Retail Load Participation in Competitive Wholesale Electricity Markets," prepared for the Edison Electric Institute and the Project for Sustainable FERC Energy Policy, January 2001.

2.1.2. Knowledgeable Consumer Response

Large industrial customers are fully aware of the value of electricity to their operations, and likely have the tools and information needed to decide how to adjust their consumption levels in response to changes in prices. Smaller customers more likely use informal rules to decide on their pattern of usage. All customer types can make use of automatic control devices to manage their energy consumption given time-varying prices.

2.1.3. Measuring Consumers' Responses

Loads are generally slower than generators in responding to prices. Few loads can respond to balancing energy prices that are known only 10 to 45 minutes in advance of each 15-minute settlement interval, nor can many loads provide responsive reserves on 10 minutes notice. Many loads cannot forecast next-day load levels for all 96 settlement intervals within the 10-15% accuracy required to avoid penalties, nor can they rapidly respond to changing market conditions (such as may be required when deciding whether to accept a curtailment).

Although these limitations imply that the services provided by loads have lower value than those provided by generators, it does not imply that the value is *qualitatively* different. The problem is *quantifying* the differences in values between the services provided by loads and generators. Indeed, the values provided by different loads are not uniform, nor are the values provided by different generators.

2.2. Market-Based Pricing and Customer Choice

Market-based pricing and customer choice provide the key market design features that connect retail and wholesale electricity markets. Market-based pricing does not require that all retail load face spot pricing. Instead, market-based pricing recognizes that retail electricity products are combinations of two services:

- the commodity product (e.g., electrical energy); and
- insurance against price uncertainty, where this insurance should be priced in an actuarially fair manner.

Customer choice then determines the market balance of demand-responsive and price-insured products.[11] With market-based pricing, customers who have more consumption flexibility and are more risk tolerant will tend to "self-insure" by choosing the demand-responsive rates. But as more customers choose demand responsive rates because of the lower expected prices, the cost of providing the price insurance will decrease due to less wholesale market price volatility. Consequently, the price-insured products become relatively more attractive. In this way, a market equilibrium will establish a balance between demand-responsive and guaranteed-price products.

[11] An example of a price-insured product is a "flat price" service under which the REP offers to serve the customer at a guaranteed constant price for all of the hours of the coming year. An REP who offers such a service may enter forward contracts to cover customers' expected loads, and will include in the guaranteed price an insurance component to cover the risk of market exposure on incremental loads.

3. TYPES OF DEMAND RESPONSE AND MARKETS FOR DEMAND RESPONSE

For purposes of this paper we define *demand response* programs in general as mechanisms for communicating prices and willingness to pay between wholesale and retail power markets, with the immediate objective of achieving load changes, particularly at times of high wholesale prices. The ultimate objective of such mechanisms is to achieve the improved market performance discussed in Section 2.

3.1. Types of Demand Response

Various types of demand response mechanisms have been proposed and/or offered in U.S. electricity markets. One logical classification scheme groups demand response activities into three generic categories: dynamic pricing (known in Texas as passive load response), load reduction programs (like interruptible service), and ancillary service programs:

- *Dynamic pricing* involves retail energy providers selling electricity at time-varying prices that reflect wholesale market costs. One common example is real-time pricing (RTP) of large commercial and industrial customers, in which hourly RTP prices reflect day-ahead or hour-ahead wholesale prices. However, new examples of residential time-of-use rates with a "critical" price component that can be transmitted on short notice are receiving considerable attention.
- *Load reduction* programs involve customers offering to reduce their electricity usage during certain time periods in return for a financial payment. The load reductions may be mandatory or voluntary, depending on the type of payment, which may be a price discount offered in advance (for mandatory programs), or a payment at the time of the reduction (for voluntary programs). Programs may be offered by incumbent utilities, load or curtailment aggregators, or ISO/RTOs. Important differences between program types suggest subdividing this category into the following three types:
  - *Load management* programs such as direct load control and interruptible/curtailable load programs can provide load relief to improve reliability during periods of low reserves, or reduce cost at times of high wholesale market costs. Consumers have traditionally been paid in advance for their *participation* in such programs, through rate discounts or monthly bill credits, rather than for their *load reduction performance*. Load reductions are typically mandatory, with substantial penalties for non-performance. Some load management programs, particularly those that are rarely operated, have actually been rate discounts in disguise. More market-based programs in the future are likely to combine pay for performance with payments for participation.[12]
  - *Energy buy-back programs* have customers agree to reduce usage in return for an incentive payment that is either established beforehand or tied to the wholesale market price. In some cases, these have been informal arrangements between

[12] Efficient interruptible programs may offer a range of participation payments (the option payment) that vary inversely with market-based payments for performance (the exercise price). That is, to receive a large discount, a customer would have to curtail load at relatively low wholesale prices and receive low payments for the load reductions.

utilities and large customers; others, such as the Energy Exchange programs operated in Oregon, consisted of formal arrangements that were activated whenever wholesale prices reached a certain level.

    - *Demand bidding programs* allow retail customers or their aggregators, such as utilities or independent REPs, to bid load decrements (*i.e*., load reductions compared to some baseline level) into day-ahead or hour-ahead wholesale energy markets.[13] The load reductions are then scheduled and dispatched in a manner similar to the scheduling and dispatch of generators. Such programs have been offered by PJM, ISO New England, the New York ISO, and the California ISO.

- *Ancillary service programs* allow customers to receive payment for providing regulation or reserve services.

Each of these categories of demand response mechanisms offers the potential to help connect the wholesale and retail energy markets. However, they differ in a number of key features, including the following:

- whether and how the amount of demand response is measured and/or validated;
- whether and how the amounts of retail demand response are taken into account in the wholesale market; and
- how the prices and/or incentive payments for demand response are set.

Under the appropriate conditions, some forms of demand response would likely evolve naturally in the development of competitive wholesale and retail markets. However, continued regulation of retail markets, and non market-based standard offer rates in markets with retail access, have been limiting the availability of dynamic pricing at the retail level. Other barriers may exist in the form of market design rules, regulatory guidelines, or lack of infrastructure (*e.g.*, advanced interval metering and communication systems).

3.2. Markets for Demand Response

Participation of demand side resources can improve the efficiency and overall performance of energy and ancillary services markets. As with some generation resources, the longer the time allowed for response, the greater the ability a demand resource has to effectively participate in a market. Consequently, demand resources are generally better suited to participate in energy markets than in ancillary services markets. However, as Hirst points out, certain demand resources (such as municipal water pumping) are actually quite well suited to provide reserve services, but "that current NERC standards prohibit the use of such pumping loads from providing this service."[14]

---

[13] The term *demand bidding* is somewhat of a misnomer in that customers are actually bidding demand *reductions*.

[14] E. Hirst, "Price Responsive Demand as Reliability Resources," http://www.ehirst.com/PDF/PRDReliability.pdf, April 2002, p. 13.

3.2.1. Energy Markets

In the other organized ISOs, demand resources have some opportunity to participate in day-ahead and hour-ahead markets. In markets without retail competition, demand resources participate in wholesale energy markets through dynamic retail rates (such as RTP), market-based interruptible programs and by special contracts with dynamic pricing characteristics. Typically participants in these programs have been large, electric-intensive industrial customers. Some of the dynamic pricing contracts have actually been developed at the behest of the customer.

3.2.2. Ancillary Services Markets

Several types of ancillary services can be provided through competitive processes. In Texas, these services are called Regulation Service – Up, Regulation Service – Down, Responsive Reserve Service, Non-Spinning Reserve Service, and Replacement Reserve Service.[15] These services are defined by the speed with which they must become available and the length of time that they must produce power when called by the system operator. The foregoing list of services is ordered with the most quickly available services first and the longest lasting services last.

Although generators can provide all of the foregoing ancillary services, loads can provide only the slower services because they tend to have neither the physical flexibility nor the communication and control technologies that are required to provide the faster services.

While demand resources may have some role in ancillary services markets, the potential is really of a second-order effect compared to the impacts demand resources can have in energy markets. Demand response in energy markets may nonetheless have beneficial indirect impacts in ancillary services markets during times of tight supply when demand response allows generation capacity to shift from energy markets to ancillary service markets. Generators will have an incentive for such a shift when demand side resources' participation in energy markets causes the price of energy to decrease relative to the price of ancillary services. Thus, encouraging efficient demand resource participation in energy markets may contribute to lower prices, reduced price volatility, and decreased market power in the ancillary services markets.

Furthermore, in a competitive market structure, the distinction between reliability and economic needs will tend to blur. As demand resources contribute to improved market performance, greater price and load stability are likely results, and the needed reserve requirement may decrease. If so, then demand response in the energy markets may indirectly reduce demand pressure in the ancillary services markets.

Encouraging efficient demand participation in energy markets should be the higher priority.

---

[15] The names of these ancillary services seem to be different in every region of the U.S. In its Order 888 on open-access transmission requirements, FERC identified a single regulation service (Regulation and Frequency Control Service), and referred to the reserve services as Operating Reserves – Spinning, Operating Reserves – Supplemental, and Backup Reserves, respectively.

3.3. Barriers to Demand Response[16]

The barriers to demand resource participation in electricity markets can be classified as customer barriers, technological barriers, regulatory barriers, and measurement barriers. Each of these categories is briefly discussed in this section.[17] The impact of load profiling on the demand response issue is also covered.

3.3.1. Customer Barriers

Customer risk aversion and inflexibility in consumption are often listed as barriers to demand participation. However, clear consistent evidence over many years indicates that some loads have some flexibility and do respond to changes in electricity prices.

Another suggested reason that customers do not accept demand responsive rates is that by-and-large retail electricity providers are not offering dynamic pricing to customers. This indeed suggests that some barriers exist, but it certainly doesn't indicate that the customer is the source of the barrier. On the contrary, several demand responsive electricity price structures have been introduced into the market in response to customer requests. Customers that have voluntarily chosen these rate structures indicate satisfaction from gaining increased control over their energy costs.

3.3.2. Technological Barriers

Technology limitations have often been a barrier to implementing more demand responsive retail pricing structures for electricity. Over time, however, this may become a smaller problem. Metering and communication advances in the past two decades have made the needed technology both more available and more affordable.[18] Nevertheless, the needed technology is not in place for the overwhelming majority of electricity consumers in the U.S. In particular, there is a lack of interval metering, a lack of advanced communication systems, and a lack of advanced systems for controlling numerous customer loads.

The technology is not in place because there are significant barriers to demand participation. In many states, there remains uncertainty over the basic issues of who owns the meter and who pays for the meter and its installation. Likewise, uncertainty about regulatory commitment to cost recovery discourages investment in the needed technology. The general absence of uniform technology standards adds a new uncertainty, worsens the impact of uncertainty about regulatory commitment, and discourages investment in the needed technology.

3.3.3. Regulatory Barriers

Regulation is often a barrier to demand-side participation in electricity markets, partly because of uncertainty about future regulatory actions. Regulatory efforts to protect consumers from price

---

[16] A review of the barriers to demand participation can be found in E. Hirst, *Barriers to Price-Responsive Demand in Wholesale Electricity Markets*, prepared for the Edison Electric Institute, June 2002.

[17] In Section 5, we discuss barriers specific to ERCOT.

[18] For example, Puget Sound has installed interval metering and provided internet-based real time consumption information for virtually every customer.

volatility play a very large role in thwarting the development of demand response initiatives. Examples of the protection include non-market based standard offer service and price caps.

Inefficient pricing of transmission and distribution services can also complicate the attainment of economically efficient demand response. These costs of providing services are largely fixed costs. If the required revenues to cover the fixed costs are collected on a volumetric basis (per kWh price), then two problems emerge related to demand response. First, the price distortion resulting from the volumetric T&D charges introduces an incentive for uneconomic by-pass in the form of both distributed generation and load curtailment. Second, in order to prevent the revenue erosion in the T&D charges, the regulator may restrict or the vertically integrated utility may not offer efficient demand response opportunities.

Finally, an ever-looming regulatory barrier is the State/Federal jurisdictional issue. Transmission and wholesale markets generally fall under the federal jurisdiction, while siting authority, distribution pricing, and retail regulation is the purview of state regulators. Nonetheless, the Federal Energy Regulatory Commission (FERC), in its Standard Market Design Notice of Proposed Rulemaking, treats demand response as a primary objective.[19] Demand response, however, ultimately comes from retail customers who are under the protection of state regulators. Although connecting retail and wholesale markets requires consistent and coordinated state and federal policies, the consistency and coordination that can be achieved remains an open question.

3.3.4. Load Profiling

Many utility rates today are based on customer-class load shapes or load profiles. The need for load profiles result from the lack of individual interval metering. While trying to best deal with the lack of individual metering, load profiling may actually hinder development of demand-responsive rates. The problem arises because, despite efforts to profile homogeneous groups of customers, some diversity within each group is bound to remain.

Furthermore, it may not be possible to accurately measure or estimate the actual load response of a load-profiled customer during critical hours. Also, it may be difficult or impossible to correctly to attribute demand response by a load profile customer to the serving REP. This likely discourages the REP from offering demand-responsive products.

Load profiling, *per se*, is not a barrier. Instead, the barrier is the lack of individual interval metering. However, both technological and regulatory barriers may inhibit individual metering for several years.[20] The result is that REPs may find it unprofitable to undertake individualized metering. Whether it would be socially beneficial for public funds to be used to install individual interval meters requires a rigorous benefit-cost analysis.

---

[19] Federal Energy Regulatory Commission, Notice of Proposed Rulemaking, *Remedying Undue Discrimination through Open Access Transmission Service and Standard Electricity Market Design*, Docket No. RM01-12-000, July 31, 2002.

[20] In Texas, the transmission and distribution utilities own the meters. Metering will not open up to competition until 2004 for commercial customers and 2005 for residential customers.

3.3.5. Quantifying the Amount of Load Response

Problems in accurately quantifying the amount of load response offered by retail customers can be a barrier to the implementation of demand response programs. The basic problem is that only energy consumption can be measured. Energy reduction must be inferred by comparing actual consumption against some baseline representing expected consumption if load response had not occurred. The baseline is typically determined by certain rules or algorithms, which introduces the possibility of gaming. Establishing the baseline and quantifying load response is particularly problematic for demand bidding programs. It is much less of an issue under traditional load management programs, as these programs have typically involved rate discounts paid in advance rather than payments for performance.[21] In contrast, dynamic pricing customers are charged for what they consume rather than being paid for how much they reduce consumption. Consequently, there is no need to measure how such customers' consumption differs from a baseline.[22]

Competitive electricity markets provide a solution to the baseline problem. If retail electricity evolves like those of other commodities, then the typical competitive retail service will consist of a forward contract for expected needs, with balancing occurring at spot market prices. The customer may also purchase (or sell) price protection through the use of call (price caps) and put (price floor) options. The forward contract in essence becomes the all-important baseline. In competition, the size and shape of the forward contract will be determined solely by the customer's hedging needs.

4. REVIEW OF DEMAND RESPONSE PROGRAMS IN OTHER ISOS

Each of the four existing FERC-regulated ISOs – the Pennsylvania-Maryland-New Jersey Interconnection (PJM), New York, New England, and California – have demand response programs in place. These programs fall into two general categories: "emergency programs" that respond to system reliability concerns; and "economic programs" that capture economic benefits according to wholesale market price levels. Unfortunately, the performance of these programs in the summer of 2001 was modest at best, and more often insignificant. Their limited performance likely resulted from the newness of the programs, their design features, and barriers to customer acceptance.[23]

We briefly report on the details and performance of these programs during the summer of 2001.[24]

---

[21] The pay in advance, not for performance characteristic of the load management programs creates serious incentive problems.

[22] It may however be very useful for a service provider to understand the aggregate demand response of dynamic pricing customers so it can make accurate price-sensitive bids into the wholesale energy market.

[23] See E. Hirst, *Barriers to Price-Responsive Demand in Wholesale Electricity Markets*, Prepared for the Edison Electric Institute, June 2002.

[24] A more detailed report on the performance of the ISO demand response programs can be found in ICF Consulting, *Policy and Technical Issues Associated with ISO Demand Response Programs*, Draft Report prepared for National Association of Regulatory Utility Commissioners, May 23, 2002.

4.1. PJM Interconnection

PJM has two demand response programs. The "Emergency Option" imposes mandatory curtailments when certain reliability conditions prevail. The participant is paid the greater of $500/MWh or their locational marginal price for the amount of the contractual curtailment. The "Economic Option" pays the participant a price equal to the excess of their locational marginal price over the standard retail price (including both generation and transmission charges).

The maximum participation on the "emergency option" in the summer of 2001 was 61.7 MW while the maximum participation on the economic offer was only 0.7 MW, yielding a combined total of just over 0.1% of the system peak.

4.2. New York ISO

The New York ISO (NYISO) also has two demand response programs. The "Emergency Demand Response Program" (EDRP) paid participants the greater of $500/MWh or the market-clearing price for loads.[25] The EDRP had an average participation of 425 MW on four occasions during the week of August 7, 2001. The maximum participation was 455 MW. This program seems to have been the most successful of all the FERC-regulated ISO demand response programs last summer.[26]

The economic demand response program, called the "Day-Ahead Demand Response Program" (DADRP), paid participants the greater of their bid prices or the market-clearing price. The participation in the DADRP program was much lower, with a maximum of 25 MW.

The combined EDRP and DADRP participation last summer was less than 0.2% of system peak load.

4.3. ISO New England

ISO New England (ISO-NE) also has two demand response programs. The emergency demand response program, called "Class 1," requires mandatory reductions on 30 minutes notice. The participant is paid based on the market clearing price for 30-minute operating reserves plus the market clearing price for the actual energy curtailed. In the summer of 2001, the Class 1 maximum participation was only 1 MW. The economic demand response program, called "Class 2," has voluntary reductions with the participant being paid the market clearing energy price for actual curtailments. The Class 2 program only goes into effect after the wholesale market price exceeds $100/MW. Last summer, Class 2 maximum participation was 20 MW. Combined maximum participation was less than 0.1% of system peak load.

4.4. California ISO

The California ISO (CAISO) had three demand response programs at the beginning of 2001. All three programs were reliability based. The "Participating Load Program," targeted toward large water pumps, allowed loads to bid into some reserve services markets. The "Demand Relief

---

[25] The NYISO wholesale market has locational marginal pricing for generation and zonal pricing for loads.

[26] Neenan Associates, *New York ISO Price-Responsive Load Program Evaluation: Executive Summary*, prepared for the New York Independent System Operator, January 15, 2002.

Program" was designed to be a last-resort to prevent blackouts. The "Discretionary Load Curtailment Program," targeted smaller industrial customers with commercial lighting and air conditioning loads.

In addition to the CAISO, other California agencies (e.g. the CPUC and the California Department of Water Resources) developed demand response programs. The multiplicity of programs and uncertainty about the financing of the programs led to confusion and probably decreased the effectiveness of the demand response initiative in California during the summer of 2001. In addition, that summer turned out to be unexpectedly mild. Nevertheless, the CAISO did have one Stage 2 System Emergency in which load reductions were obtained by the Demand Relief Program (162 MW) and the Discretionary Load Curtailment Program (22 MW).

4.5. Other Demand Response Initiatives

Significantly more load reduction was achieved during the summer of 2001 through programs operated by the local equivalents of Qualified Scheduling Entities (QSEs) and REPs than through the ISO demand response programs. These QSE and REP programs notably included interruptible programs and demand buy-back programs. In PJM, almost 1,800 MW of load reduction was achieved from interruptible programs in high-priced hours. Likewise, in California, 760 MW of load reduction was achieved via interruptible programs operated by the QSEs and REPs. Prior to FERC's imposition of the West-wide wholesale price cap, Portland General Electric had a very active internet-based demand buy-back program.

The real-time pricing programs operated by Georgia Power and Duke Power indicate significant demand response when prices hit or exceed the $350 – $500/MWh level. For example, in August 1999 when Georgia Power's real-time prices exceeded $1,000/MWh, customers responded by reducing load by about 800 MW (out of a total of about 5,000 MW participating on RTP) or about 20%.[27] While prices in the Southeast never reached that level in 2001, the 1999 Georgia Power maximum RTP demand response was greater than the combined maximum demand response from all the ISO demand response programs in 2001.

5. DEMAND PARTICIPATION IN ERCOT'S MARKETS

ERCOT is a single control area with 70,000 MW of generation capacity, 37,000 miles of transmission lines, and a peak demand of 57,600 MW. ERCOT is organized as an independent, not-for-profit organization with a stakeholder board. Because ERCOT is entirely within the State of Texas, it falls under the regulatory jurisdiction of the Texas PUC.

The *qualified scheduling entity* (QSE) is the business entity that directly interacts with ERCOT. *Retail electric providers* (REPs) provide electricity services to retail customers. REPs include competitive retail providers as well as municipals and coops. All REPs schedule through a QSE. The QSE must self-provide all wholesale energy to serve their load – i.e. the QSE must have generation resources and/or enter into bilateral contracts with generators to cover load obligations and must submit a balanced schedule to ERCOT. The QSE may self-provide ancillary services other than replacement reserves and balancing energy, but this is not required.

---

[27] S. Braithwait and M. O'Sheasy, Customer Response to Market Prices – How Much Can You Get When You Need It Most?" *EPRI International Pricing Conference 2000*, Washington, DC, July 2000.

ERCOT operates an ancillary services market an ancillary services market for QSEs that do not self-provide and a balancing energy market for all participants. ERCOT settles all wholesale energy accounting between QSEs. The QSE then settles accounts with the REPs. The retail customer interacts only with the REP and is thusly separated from the wholesale market by two intermediary entities – the REP and the QSE.

ERCOT does not operate an energy market, other than the real-time balancing energy market. ERCOT does operate ancillary service markets for regulation, responsive spinning reserve, non-spinning reserve, and replacement reserve. Ancillary service providers must be certified by ERCOT.

ERCOT believes that the balanced schedule requirement is essential to maintain system reliability. ERCOT is currently examining the protocol changes that would allow relaxation of this requirement and thereby allow market participants to lean on the balancing energy market. It is anticipated that a relaxed balance schedule pilot will be put in place on October 1, 2002. Nonetheless, there are currently no plans to allow unlimited imbalances in schedules, as this would require fundamental changes to market protocols and a new role for the ERCOT ISO as an energy pool operator. ERCOT stakeholders believe that the reliance on bilateral contracts rather than a centralized energy market, and the relative smallness of the balancing energy market, are desirable features that improve the performance and reduce price volatility in ERCOT markets.[28]

5.1. Load Participation Opportunities in ERCOT

Demand resources have limited opportunities to participate in wholesale markets in ERCOT. Nominally, demand resources can participate in the balancing energy market and in the reserve services markets. Demand resources can participate in three ways: via *passive load response*, as a certified *balancing up load (BUL) provider*, and as a *load acting as a resource (LaaR)*. The actual opportunities for demand participation in ERCOT currently appear to be only in the ancillary services markets. This is problematic because energy markets provide the greatest potential source of benefits from demand participation. Unfortunately, the current Texas market structure is unable to achieve these benefits because of the severely limited opportunities for demand participation in energy markets.

5.1.1. Passive Load Response

Passive load response involves a customer responding to a price signal to change consumption. As described in Section 3.1, so-called passive load response involves retail electric providers selling electricity at time-varying prices that reflect wholesale market costs. However, in the ERCOT markets, passive load response occurs formally only in the balancing energy market, and it is the QSE that must "deliver" that passive response into the ERCOT market. However, it is the customer that must provide curtailment in response to a price signal. Passive response in ERCOT markets relies on the development of incentives for the appropriate price signals to be sent beyond the QSE through the REP to the customer. As discussed below, the absence of a centralized day-ahead market hinders the effectiveness of the conduit conveying passive load response from the customer to ERCOT.

[28] See S.R. Jones, ERCOT CEO, "The New Texas Wholesale/Retail Market," presentation to the Federal Regulatory Energy Commission, January 23, 2002 (especially Slide 10).

5.1.2. *Certified Balancing Up Load Resource*

In the Balancing Up Load (BUL) market, curtailable loads will be paid market-based real-time energy and capacity prices. Loads can bid into the balancing energy market. When a customer's bid is accepted, the customer receives the marginal clearing price of energy as well as a capacity payment that is equal to the marginal clearing price of capacity in the non-spinning reserves market. However, the load must be certified as meeting response requirements and must have both a REP and QSE that are amenable to delivering load into the BUL market.[29]

5.1.3. *Load as a Resource*

Some demand resources can participate in the ERCOT reserve services markets. The ERCOT certification of LaaRs requires greater performance standards than required in BUL markets. In addition, there are also greater metering requirements (telemetry). The LaaRs that meet the performance qualifications are typically large industrial loads with automatic control technology. A requirement of LaaRs is that they have to agree to be available for curtailment by the ERCOT ISO in an emergency situation. LaaR participation therefore may increase system reliability.

5.2. Current Extent of Load Participation in ERCOT Markets

Passive load is neither registered nor certified by ERCOT. Consequently, it is difficult to assess exactly how much passive load response actually exists. However, it is generally perceived that there is little passive load response in ERCOT markets. It is a central objective of the PUC to identify barriers to passive load response and to take appropriate actions to increase passive demand response.

As of August 2002, the BUL market was not yet operating.[30] Currently, the main opportunity for demand-side participation in ERCOT markets is for LaaRs to provide ancillary services in the day-ahead ancillary services market. Since the market opened on January 1, 2002, three customers have been certified as LaaRs, for a total of 503 MW. Three more customers have applied for certification, for a total of an additional 547 MW. These customers will be represented by three different QSEs. When these additional customers receive certification, a total of 1,050 MW will be available to the ISO for curtailment in emergency situations.[31]

Before the retail market opened, there were approximately 3,500 MW of load in ERCOT on interruptible tariffs. However, the old programs no longer exist; so these interruptible loads are no longer available to respond to market conditions.[32] Thus, it appears that the Texas retail competition design dramatically decreased the amount of passive load response, at least temporarily.

---

[29] Qualified customers of all load-serving entities can potentially participate in the BUL and LaaR markets, even if they are customers of municipally owned utilities and cooperatives that have not opted into retail competition.

[30] Based on conversations with PUC staff.

[31] Based on conversations with PUC staff.

[32] Based on conversations with PUC staff.

5.3. Barriers to Demand Participation in ERCOT Markets[33]

This review has discovered several factors that appear to limit demand participation in wholesale markets in Texas. We have organized these barriers into three prioritized groups. The most critical group, *Major Barriers to Demand Participation*, identifies the barriers that have kept meaningful demand response participation opportunities from developing. The second grouping, *Impediments Limiting Demand Participation*, identifies factors that limit participation in the current opportunities. The third grouping, *Potential Barriers*, identifies factors that might limit future demand responsiveness, but that currently are not limiting because of the effect of the barriers in the first two groups.

It is worthwhile mentioning that because ERCOT is completely under the regulatory jurisdiction of the Texas PUC, there is not the multi-jurisdictional complication that faces other ISOs attempting to encourage demand-side participation in wholesale markets. Thus, while there are several significant barriers to demand response in ERCOT markets, the avoidance of the multi-jurisdictional complexity uniquely positions Texas to address these barriers, to achieve market design reforms and to facilitate demand-side participation.

This section identifies the barriers to demand participation in ERCOT markets. Section 6 provides more detail on these barriers, and Section 7 makes recommendations for market rule changes to address these barriers.

5.3.1. Major Barriers to Demand Participation

The single largest barrier to demand-side participation in ERCOT is the absence of centralized day-ahead and same-day energy markets.

Other contributing factors include:

- the balanced schedule requirement; and
- the degrees of separation between the customer loads and the wholesale markets.

The combination of the balanced schedule requirement and the *complete* reliance on bilateral contracts may create significant transaction costs associated with recruiting demand responsive load, especially passive load response. That is, accepting some demand response may cause the QSE to have to re-balance its bilateral contracts. Centralized energy markets could greatly reduce the transaction costs, facilitate demand participation and increase efficiency overall.

Bilateral contracts are a natural part of any well functioning electricity market. Market participants ordinarily want to hedge themselves against uncertain day-ahead and same-day market prices. Consequently, they use bilateral contracts to arrange in advance for much of their load requirements. However, relaxing the balanced schedule requirement and creating a centralized energy market would give participants the flexibility to go to the market for deviations from their contracted loads rather than having to make short-notice adjustments to bilateral contracts. The increased flexibility would facilitate demand participation, especially passive load response.

---

[33] Many, but not all, of these barriers are identified and discussed in B. Begg and J. Zarnikau, "Task Force on Demand-Side Resources and Demand Responsiveness: Status Report," January 2002.

The end result of the current market rules is that the QSEs and REPs have little incentive to offer demand resource programs or dynamic pricing products, and customers have little opportunity to respond to wholesale market conditions. Furthermore, customers are dependent upon the QSE and their REP to provide access to whatever opportunity there may be.

Why then don't private centralized markets develop? Possible explanations could be that organizing and operating a centralized energy market is a "public good" and thus subject to extreme free riding by participants.[34] Also, the uncertainties involved with establishing a brand new market in a traditionally regulated industry also likely discourage private investment the organization of the market.[35] APX, a private enterprise that has organized electricity markets in California, reported that they had tried to organize a centralized market spot market in Texas, but found insufficient interest because of the practice of bilateral contracting.[36] In a recent interview with PUC staff, Green Mountain Energy stated that a centralized market is not needed, but that a centralized market would increase price transparency and that Green Mountain would participate in such a market if it existed.[37]

Addressing these major barriers would greatly increase the opportunities for demand to participate in energy markets and would yield the greatest benefits in terms of economic efficiency and market performance. This should be a high priority.

5.3.2. Impediments Limiting Demand Participation

This second grouping identifies barriers that hinder demand participation in the limited opportunities that currently exist. These impediments are:

- restrictive and complicated certification process for BULs and LaaRs;
- difficulties in accurately measuring load response; and
- lack of interval metering and reliance on load profiling, which together create disincentives for offering demand response opportunities.

Both the length and the density of the ERCOT Section 6 protocols on ancillary services attest to the complexity of the settlement and certification processes for BULs and LaaRs.[38] The problems associated with measurement and load profiling have been discussed in other sections of this report.

---

[34] Production of public goods typically involves high fixed costs and near-zero marginal or variable costs. A public good has two essential characteristics – non-rivalrous and non-excludable consumption. The free rider problem arises as each potential market user waits for someone else to provide the good.

[35] The California experience and subsequent retreat from deregulation along with the Enron collapse and other questionable energy trading activities have undoubtedly caused potential investors to question the strength of the commitment to wholesale and retail electricity competition.

[36] Reported at the DSWG meeting of July 8, 2002.

[37] Based on conversations with PUC staff.

[38] The difficulty the DSWG has had in writing and getting agreement on the "BUL for Dummies" and "LaaR for Dummies" documents is further evidence.

Addressing these impediments will improve the ability of loads to participate in the ancillary services and BUL markets. However, addressing the impediments should be a second priority to the addressing the major barriers.

5.3.3. Potential Barriers

This third group has the lowest priority as far as the need for immediate attention. Potential future barriers to demand resource participation in Texas energy markets are:

- non-market based standard service offers (e.g., the "price-to-beat"); and
- inefficient transmission and distribution pricing.

Non-market based standard service offers can undermine the effort to achieve market-based pricing and customer choice. As discussed in Section 2, market-based pricing and customer choice are the key design features that connect the wholesale and retail markets. Fortunately, the impact of non-market based standard service offers is somewhat limited, as Texas customers larger than 1 MW do not have the price-to-beat protection.

As discussed in Section 3, inefficient transmission and distribution pricing can hinder efficient demand response for several reasons. First, the current pricing of transmission may cause too much load to be shifted away from particular hours of the four summer months. That is, the relative price of transmission in those particular hours to the price of transmission in other hours is likely to overstate the cost differential. Second, transmission prices are not dynamic, and therefore cannot accurately reflect rapidly changing transmission constraints. Third, because transmission and distribution pricing is largely volumetric, but transmission and distribution costs are largely fixed, efficient demand response may cause revenue collection problems for transmission and distribution utilities. As a result, there may be resistance to the QSEs and REPs encouraging economically efficient demand response.

It is important to recognize these potential barriers, while also recognizing they are not the chief culprits currently discouraging demand participation.

6. EFFECTS OF EXISTING ERCOT MARKET RULES ON EFFICIENT DEMAND RESPONSE

The purpose of encouraging demand response to market conditions is to find ways of maximizing the net benefits that consumers derive from electricity. As previously mentioned, achieving efficient demand response requires a process that involves three broad steps:

1. Market conditions are communicated to consumers through either a price signal or a quantity (e.g., curtailment) signal.
2. Consumers respond to the signal that they receive.
3. The system operator measures consumers' responses and pays for these responses through the settlement process.

Mirroring these three steps, efficient demand response is inhibited by factors that:

1. give consumers inaccurate signals about market conditions;
2. inhibit consumers from efficiently responding to the signals that they receive; or

3. inaccurately measure and reward consumers' responses.

This section discusses existing characteristics of ERCOT's markets that may create each of the foregoing types of factors. The subsections are ordered accordingly.

6.1. Signaling Consumers About Market Conditions

6.1.1. Energy Markets

Because of the role that the price mechanism plays in communicating information about market conditions and in inducing efficient response to these conditions, a primary consideration in evaluating the ERCOT's operations and market design is the extent to which they convey accurate and timely price signals to generators and (through REPs) to customers.

ERCOT's lack of centralized day-ahead and hour-ahead energy markets makes price information costly. The lack of firm, transparent day-ahead and hour-ahead energy prices is likely to increase the costs to REPs of offering to their customers products based upon dynamic power system and market conditions. This barrier limits the participation of both passive and active load response in ERCOT.

Regarding the accuracy of the price mechanism, wholesale prices in ERCOT are zonal, which means that the price at each location is an average price for the zone, rather than the (higher or lower) efficient locational price. Zonal prices therefore provide an inaccurate signal to energy providers about the incremental costs of serving loads at various locations.

6.1.2. Ancillary Service Markets

Loads acting as resources in ERCOT are paid market-clearing prices for responsive reserves, non-spinning reserves, replacement reserves, and Balancing Up Load (BUL). To the extent that loads and generators provide the same service, this is a good thing: as a matter of efficiency, all resources, whether loads or generators, should receive the same prices for each service. An exception may be made if, for example, load provide environmental benefits relative to generators, in which case some non-market mechanism might be developed for the purpose of paying loads an extra amount for these benefits.

In ERCOT, the efficiency of demand response may also be compromised by the manner in which transmission costs are recovered. Because these costs are recovered through charges on customers' coincident peak demand during the four summer months, the costs that customers are willing to incur in attempting shift loads away from the four peaks may significantly exceed any possible resulting reduction in power system costs: even if transmission costs are not completely fixed, they are may not be as dependent on peak loads as implied by the method of cost recovery.[39]

[39] In England, a very similar system has given rise to a mini-industry of peak-load prediction services, so that customers will have information that might allow them to shift loads away from hours that are predicted to be the peak hours. Both the prediction services and the load shifting are a complete waste of society's resources, as they have no effect on transmission costs, but instead affect only the allocation of these costs among customers.

6.2. Consumers' Responses to Market Signals

Consumers may lack information or understanding of programs, including market rules, settlement procedures, or program designs. They may also misperceive the financial effects of penalties for non-performance. Furthermore, ERCOT's administration of markets through REPs and QSEs, including the requirements for bilateral contracts and balanced schedules, seems to inhibit demand response in the following ways:

- The standard contracts offered by most REPs include "bandwidths" to discourage deviations from historical consumption patterns.
- The requirement that each QSE have a balanced schedule discourages demand response by penalizing (or threatening to penalize) demand responses that result in imbalances, even when a particular QSE's imbalance might help system reliability. The balanced schedule requirement has led to proposals to distinguish "legitimate" price responses from gamed load forecasts that enable QSEs to reap profits from intentional imbalances, even though the reliability effects of the resulting imbalances may be identical for the "legitimate" and "illegitimate" response, and even though reliability might be improved by a particular QSE's imbalance.[40]
- A QSE may violate the Protocols by providing more than 120% of its responsive reserve commitment.[41] This provision of the Protocols can cause QSEs that have inexpensive

[40] At least one news report and three documents refer to the problems cited in the text. *Utilities Biweekly Report* of April 2, 2002 (http://www.tdpud.org/index.cfm) states, "The Texas Public Utility Commission (PUC) alleges that six companies abused Texas' newly deregulated electricity market and profited by manipulating the wholesale energy market… The PUC's Market Oversight Division reports that a quirk in the deregulated system allowed the QSEs to overschedule demand last August and charge high prices to companies needing additional power."

ERCOT Working Group on Demand Side Resources and Demand Responsiveness, *Meeting Notes: February 25, 2002*, March 3, 2002, p. 1, refers to "a proposal by the staff's consultant, Dr. Shmuel Oren, to address gaming opportunities caused by QSEs that intentionally manipulate their load forecasts to sell balancing energy during periods of anticipated high prices. Under the proposal, a QSE could receive a credit or pay a penalty any time its load level changed."

Frontier Associates, "Comments on Interim Memorandum from Christensen Associates," July 12, 2002, p. 2, states "Professor Oren has offered a proposal intended to reward 'legitimate' price chasing by loads while discouraging the gaming of load forecasts in order to reap profits from the balancing energy market through an intentional imbalance."

J. Zarnakau, *Transitioning ERCOT's Demand-Side Resources into the New Market: A Scorecard*, December 2001, slide 10, states "…some scheduling entities have (allegedly) figured out how to game the system by over-scheduling loads…"

[41] At least three documents refer to the problem cited in the text. Frontier Associates, "Comments on Interim Memorandum from Christensen Associates," July 12, 2002, p. 3, states "A particular QSE could provide more than 120% of its responsive reserve commitment during a deployment, thus violating the Protocols."

ERCOT Working Group on Demand Side Resources and Demand Responsiveness, *Meeting Notes: February 25, 2002*, March 3, 2002, p. 2, states "…under the present Protocols, if a LaaR bid conservatively and ended up providing more than 120% of its bid amount, the LaaR could conceivably be penalized for over-providing the reserves."

ERCOT Task Force on Demand Side Resources and Demand Responsiveness, *Meeting Notes: August 29th, 2001*, August 30, 2001, p. 4, reports initial committee approval of the following proposal: "But in no instance, the resource would be paid for more than 120% of the quantity reflected in their offer."

responsive reserve resources (including responsive loads) to leave the resources unused while other QSEs use more expensive resources.

- As a result of the Protocols, large QSEs may have greater abilities and incentives to deliver LaaRs to ERCOT. Thus, the rules may arbitrarily improve the competitive position of large QSEs and limit the LaaRs opportunities of REPs and loads served by smaller QSEs.

6.3. Measuring the Services Provided by Consumers

ERCOT's debate about the treatment of loads as resources has been confounded by ERCOT's lack of methods for measuring the quantities of services provided by generators or loads. An efficient treatment of resources would recognize that they can provide a spectrum of values. For example, a generator that is 98% reliable in providing 10 MW of responsive reserves is clearly more valuable than a generator that provides 10 MW with only 90% availability; but the latter generator nonetheless offers a valuable service. Similarly, a consumer who promises 10 MW of load reduction and delivers those 10 MW when demanded is providing a more valuable service than one who promises 10 MW but delivers only 6 MW; but the latter consumer is still providing a valuable service.

The difficulty of recognizing that resources provide a spectrum of values is the fundamental cause of many of ERCOT's debates about the treatment of loads as resources. For example, it leads to questions about whether loads should be treated the same as or differently from generation resources, and about whether more stable and predictable loads should be allowed to offer certain services while less stable and predictable loads are not allowed to offer these services. These questions (and others like them) can be resolved by creating service measure methods that recognize the value of a resource depends upon its performance and predictability, characteristics that can be quantified and applied without discrimination to all resources, regardless of type.[42]

Given the lack of hourly metering for mass-market customers, restructured power markets have adopted different versions of load profiling to allocate consumers' metered monthly usage to hours for purposes of charging their energy supplier. Because of diversity in energy usage patterns within load profile segments, customers with low load factors are lumped together with customers with high load factors; and customers with relatively high on-peak use are lumped together with customers with relatively high off-peak use. The barrier to greater demand response, however, is the lack of interval metering, rather than the load profiling process itself. With respect to measuring demand response, the fatal limitation of load profiling is that it cannot indicate how individual customers respond to market signals. Because each load-profiled customer would know that its credit for demand response depends upon the average response of its customer segment rather than upon its own individual response, each customer will lack any financial incentive to respond.

---

[42] Methods for creating such service measures are beyond the scope of this paper. The interested reader may, however, infer a general method from L.D. Kirsch and R. Rajaraman, "Profiting from Operating Reserves," *The Electricity Journal*, 11(2): 40-49, March 1998.

7. RECOMMENDATIONS

Our recommendations reflect our belief that demand response programs should provide the greatest possible net benefits to electricity consumers. Our recommendations therefore attempt to subject demand response programs to legitimate benefit-cost tests[43] that are market-driven to the extent feasible.

In the long run, efficient demand response requires efficient wholesale electricity pricing. Efficient wholesale electricity prices not only indicate the value of demand response but also provide a means of financing demand response. For example, when demand response is most valuable in the Dallas load pocket, high locational wholesale prices in that load pocket quantify the market-based value of demand response. Furthermore, an REP can profitably create a demand response program by which it pays consumers for load reduction in the load pocket because the REP will be able to sell that load reduction to the wholesale market at the high locational price. But note that high locational *wholesale* prices do not require high locational *retail* prices. High locational wholesale prices within load pockets can make locational demand response programs profitable for REPs even in the absence of locational pricing of retail loads.

In the absence of efficient wholesale electricity prices, it is more difficult to create efficient demand response programs. There are two particular difficulties. First, to determine the benefits of demand response, it is necessary to somehow estimate the locational value of demand response. For example, we know that demand response in the Dallas load pocket is often more valuable than demand response elsewhere in the power system; but without efficient wholesale prices, we need some non-market means of determining what that value might be. Without knowing the value of demand response, we cannot distinguish a demand response program that creates wealth from one that destroys wealth.

Second, in the absence of efficient wholesale prices, some administrative method must be developed for financing demand response. Without efficient locational wholesale prices, an REP will not have a market-based incentive for paying consumers in load pockets to reduce load because it will not be able to sell load reductions to the wholesale market at high locational prices. To create efficient demand response programs, it will be necessary for ERCOT or some government agency to pay consumers for the load reductions; and ERCOT or the government agency will then need to recover the expended funds through a charge on other electricity consumers or from revenues collected from Texas taxpayers.

There is, of course, the possibility of creating demand response programs for which benefits are vaguely defined and vaguely considered. For example, it might be decided that the Dallas load pocket needs 50 MW of available load reduction during peak periods for the sake of providing some unquantified "reliability benefits." At the very least, such programs should aim to achieve their goals at least cost.

Because Texas will not be capable of implementing efficient market-based pricing for several years, our recommendations are divided into three parts: those that can be implemented by the summer of 2003; those that can be implemented within a couple of years or so; and those that can be implemented only over a longer time frame. We note that many of our recommendations are consistent with FERC's proposed Standard Market Design. We emphasize that our

---

[43] We note that some states have tests that double count benefits and/or fail to count some costs.

recommendations generally apply to all loads and generation resources, regardless of size, though we recognize that the relatively high per-MWh cost of administering demand response programs to smaller customers often makes it cost ineffective to do so.

The following recommendations generally apply to all loads and generation resources, regardless of size. We believe that demand participation should be extended to all classes on a non-discriminatory basis, but note that the per-MWh cost of administering demand response programs to smaller customers often makes it cost ineffective to do so.

7.1. Short-Term Recommendations

7.1.1. Select a Standard Method for Measuring the "Baseline Loads" of Curtailed Customers

Curtailment programs need reasonably accurate measures of the quantities of load curtailed, which essentially means estimating the "baseline load" that each curtailed customer would have consumed in the absence of curtailment. Because all ERCOT loads are participating in the same market and because the value of curtailed load is independent of the QSE or REP that serves a curtailed load, a standard method for measuring baseline loads should applied to all curtailable customers in ERCOT.

There are several candidate methods for estimating baseline loads. One such method would set the baseline load for curtailed hours equal to the load level in the hour(s) immediately preceding curtailment. Another method would set baseline loads according to the loads in similar hours of recent days of the same day-type. Still another method – which is the one that we prefer – would use statistical analysis to examine how loads before and after a non-curtailment period can best "predict" the loads for that non-curtailment period; and the predictions of this analysis, having been validated for non-curtailment periods, can be used to "predict" the baseline loads for curtailment periods.

Texas has considered a variety of such methods. The choice of methods needs to be brought to a resolution based upon analysis rather than guesswork. Texas should therefore undertake a modest-sized study that compares the accuracy of different methods of predicting baseline loads; and based upon the results of this study, Texas should pick a method that can be uniformly applied to all curtailable customers.

As noted in Section 3, well-functioning competitive electricity markets provide a long-run solution to the baseline problem. In a fully competitive market, forward contracts would provide clearly defined property rights to electricity and would establish the baseline from which to measure curtailments. In markets in which customers can choose to continue to purchase electricity under guaranteed price contracts without quantity restrictions, the additional requirement (and inaccuracy) of estimating a baseline load will result in smaller curtailment payments.

7.1.2. Survey Customers, REPs and QSEs on BUL and LaaR on Their Demand Response Program Participation Decisions

BUL and LaaRs are the only explicit opportunities for load participation in ERCOT markets. The BUL market is not yet operating and only three customers are currently certified as LaaRs. It would be useful to gain insights as to why customers are not participating to a greater extent and what incentives and disincentives are perceived by the QSEs and REPs in recruiting customers for participation. Similarly, it would be useful to survey customers on their desire for passive load response opportunities such as real-time pricing and critical price time-of-use pricing, and to survey REPs on their perceived incentives and disincentives for offering dynamic pricing products to retail customers. It might be particularly useful to survey those retail customers who were on dynamic pricing and interruptible rates prior to markets opening.

7.1.3. Develop Pilot Curtailment Programs

We suggest that Texas develop *bid-based* curtailment programs for those load pockets in which it believes it has the greatest need for load relief. We suggest that, for pilot programs in the summer of 2003, these programs be structured in either of the following two ways.

The first type of program is similar to BULs and to the emergency demand response programs found in the New York, New England, and PJM ISOs. It would work as follows. Prior to the summer season, QSEs or REPs would ask customers to name the quantities of load they are willing to curtail (or their firm power levels), the prices at which they are willing to accept curtailments, and the maximum duration and number of curtailments that they are willing to accept during the forthcoming summer season. ERCOT would then have the right to curtail customers under the conditions named by the customer. ERCOT would pay customers for the reserves that they provide and for actual curtailments, and would pay each customer a curtailment price at least equal to its offer price.[44] QSEs and REPs would be paid fees that are related to the payments made to the participating customers who they serve.[45] When ERCOT needs to curtail customers, it would accept curtailments in ascending order of the customers' bid prices (adjusted to consider the non-price components of the bids). It would be most efficient for ERCOT to determine the volume of curtailments by comparing the curtailment bid prices to the costs of incremental generation resources; but the volume of curtailments might also be determined according to operators' MW needs.

The second type of program is similar to the economic demand response programs found in the New York, New England, and PJM ISOs. QSEs or REPs would again arrange in advance for customer participation in the program and would again be paid fees that are related to the payments made to their participating customers. In this case, however, customer participation would not require that the customer accept curtailments. Instead, in each load pocket, ERCOT would offer to pay all curtailable customers a curtailment price of ERCOT's choosing, applicable to a period beginning at least (say) 90 minutes after the offer is made and lasting at least (say) two hours. This curtailment price could be uniform over all hours of the curtailment or it could change each hour; but either way, prices would be announced in advance. Based upon the prices offered, the amount of advance notice, and the possible length of curtailment, each customer

---

[44] To create an efficient auction, it would probably be best for all curtailed customers within a load pocket to receive the highest offer price actually accepted for curtailment. A less efficient alternative would pay each curtailed customer its own offer price. This latter alternative is less efficient because it would induce gaming of bids: each customer would have an incentive to guess how far up the bid stack the curtailments might go; so customers who have relatively low curtailment costs might bid too high, resulting in curtailments of customers with relatively high curtailment costs.

[45] These fees may be per-MW amounts, per-MWh amounts, and/or percentages of the payments for reserves and curtailments.

would decide what portion (if any) of its load it was willing to curtail in each hour of the curtailment period. Depending upon how the total customer acceptances compare to the power system's need for load relief, ERCOT would then decide which offers to accept.

Under both types of programs, baselines would be established as described in Section 7.1.1. Customers would be penalized for failure to perform. ERCOT would need to somehow recover its curtailment costs from customers, which would most efficiently be accomplished by recovering the costs of curtailment in each load pocket from the loads in that load pocket rather than from all ERCOT loads.

The two types of programs share some important strengths, all of which arise from their being market-based. Because they are voluntary, they both create *expected* economic benefits for the customer. Because they only reward performance rather than participation, they both assure that ERCOT gets the curtailments for which it pays. And both approaches allow ERCOT to balance the real-time costs of demand resources with the real-time costs of supply resources – if ERCOT has reasonable estimates of real-time supply costs.

The two types of programs also have important differences. The first approach would give ERCOT advance information about the curtailable load to which it has rights, and allows ERCOT to decide which of those rights it will exercise in real time. The second approach leaves the last-minute choices up to the customer, does not assure ERCOT that it will get the volume of curtailments that it might want, and does not provide ERCOT with information about the different costs that customers incur when they are curtailed. The first approach is therefore better from the standpoints of system operator control and of efficient dispatch of demand resources.

7.1.4. Develop Benchmarks of the Benefits of Demand Response Programs

To distinguish demand response programs that create wealth from those that destroy wealth, it is necessary to estimate the costs and benefits of these programs. The cost information is relatively easy to develop: the costs of metering and communication equipment are available from the vendors of that equipment; and the costs to customers can be inferred from bidding procedures such as those outlined in Section 7.1.2. The greater challenge is estimating benefits.

In a power system with accurate price signals, the current benefits of demand response programs can be measured by the prices of the energy and ancillary services that demand response provides, and the future benefits can be measured according to forecasts of these prices. For Texas, it is instead necessary to measure benefits according to estimates and forecasts of what prices would be if the market were designed to provide accurate price signals. This basically requires methods for estimating and forecasting the locational marginal costs of energy and ancillary services.

Because efficient prices should approximate marginal costs, we recommend that Texas develop marginal cost-based benchmarks of the benefits of demand response programs. There already exist models that are capable of making the necessary estimates and forecasts. Based upon a common set of modeling and input assumptions, consistent benchmarks can be applied to all demand response programs. These benchmarks would be differentiated by the locations and time periods of each demand response program.

7.1.5. Evaluate Metering Policy

To reward a customer for its demand response, it is necessary to have accurate measures of that customer's loads over short periods such as an hour or fifteen minutes. No amount of load profiling or guessing will do the job: only individual customer metering can provide the needed accuracy.

We therefore suggest that metering policy be re-evaluated in light of Texas' significant market changes. In particular, benefit-cost analysis can indicate the customer types and conditions under which expanded interval metering of individual customers is warranted.[46] A separate benefit-cost analysis can indicate the conditions under which meters should be installed for the purpose of sampling customer loads. Texas may have additional purposes to which a metering study should be applied.[47]

7.1.6. Improve the BUL and LaaRs Document

Despite considerable voluntary effort, understanding the settlement and certification processes for BULs and LaaRs remains a challenge. Given that BUL and LaaR participation is (thus far) the main mechanism for incorporating demand response into ERCOT markets in the short to intermediate term, it would be useful for ERCOT to dedicate some staff resources to reviewing and revising the protocols regarding BULs and LaaRs. This review should look for ways to simplify the certification process and to make the settlement process easier to understand.[48]

7.1.7. Require Each Resource to Pay Its Own Overhead Costs

Each load or generator resource that wants to participate in ERCOT's markets should pay for the metering, communication, and control costs necessary to allow their market participation. If each resource thus pays the overhead costs of its participation, then ERCOT's markets should be open to resources of all sizes. The fact that these costs are relatively high for small consumers means that it is generally not efficient for such consumers to participate in certain markets; but if some small consumers are willing to pay these costs, their participation will not impose costs on other market participants.[49]

7.1.8. Evaluate the Benefits and Costs of Our Long-Term Recommendations

In Section 7.3 of this report, we suggest that Texas adopt an efficient pricing system that includes centralized day-ahead and real-time markets, three-part bidding, and locational pricing. For

---

[46] In general, metering becomes more valuable as price volatility rises, as customer responsiveness increases, as customer size increases, and as metering costs fall.

[47] In the market for Responsive Reserves, ERCOT uses telemetry to get information on every generator and every certified load and has access to actual performance every 10 seconds. This may satisfy the need for metering loads that participate in the LaaRs program. Metering of participation in the BULs program is more of a problem because telemetry in not required for BULs loads.

[48] The "BUL for Dummies" and "LaaR for Dummies" documents produced by the DSWG would provide good starting points for ERCOT staff.

[49] We do not know the extent to which current policy may already require that those wanting to participate in ERCOT markets pay metering, communication, and control costs.

numerous reasons mentioned throughout this report, we strongly believe that such a pricing system is very important for inducing efficient demand response to market conditions. Furthermore, based upon the experience of America's other ISOs, we expect that the benefits of such a pricing system will outweigh its costs.

Nonetheless, we understand that some parties wonder if the benefits of these significant changes in market rules will in fact exceed the costs. We therefore suggest that Texas may wish to conduct a study that examines this question. Part of this study could address the comparative net benefits of ERCOT organized and operated energy markets versus private organization and operation.

7.2. Intermediate-Term Recommendations

7.2.1. Develop Non-Discriminatory Measures of Performance

To assure least-cost provision of energy and ancillary services, ERCOT needs an objective method for quantifying the relative values of resources according to their performance; and it therefore needs to reward resources according to the *effective quantities* of services that they provide. *All* resources – generators and loads – should have their *effective quantities* measured according to exactly the same rules. *All* resources should face the same price per *effective quantity* supplied.

The method for measuring *effective quantities* should be able to quantify the uncertainty in each resource's performance.[50]

- The same method should apply to both load and generators because, for both loads and generators, uncertainty reduces the value of the service provided. Just as loads are not able to precisely predict their next-day levels, so generators' are not able to precisely predict their next-day availability. In both cases, the imprecision leads to errors in the provision of ancillary services.
- The services promised by more predictable loads are more valuable than those of less predictable loads. Adjustment for performance would allow comparable measurement of the services provided by these two types of loads, thereby allowing them to compete with one another.
- For each load (or aggregation of loads) and each generator, historic performance should provide a reasonable basis for assessing this uncertainty.

The methods for measuring *effective quantities* should consider the differing operational flexibility of different resources. Some load and generation resources can respond in fixed MW amounts, while others are more flexible. Some resources can follow system operator dispatch instructions more quickly than others. All other things equal, the more flexible resources are more valuable; but the less flexible resources still have value.

The methods for measuring *effective quantities* should be blind to aggregation. The quantity of a service that the power system receives from any resource is, as a physical fact, the same

[50] Methods for measuring effective quantities are beyond the scope of this paper. For insights regarding what these methods might be, see L.D. Kirsch and R. Rajaraman, *op. cit*.

regardless of the identity of the QSE from whom the resource receives service. The measured service provided by a particular resource should therefore be the same regardless of the QSE that serves the resource. If the measurement of services is not blind to aggregation, then the performance of any individual resource will be measured differently depending upon the resource's choice of QSE. In Texas, such measurement differences occur, for example, because the performance of each QSE is measured (in part) according to the QSE's imbalances as a *percentage* of load; so a highly variable 50 MW load might create penalties for a small QSE but not for a large QSE, even though the adverse effects of that load on system frequency would be identical regardless of its choice of QSE. Such differences in measurement are capricious because the resource's contribution to system costs and performance does not depend upon its choice of QSE. In this example, it would be more efficient for ERCOT to assess each QSE's performance in MWs, rather than as a percentage of QSE load, because costs depend upon MW uncertainty. Judging performance as a percentage of QSE load gives an arbitrary competitive advantage to large QSEs and ultimately discourages competition by promoting larger and therefore fewer QSEs.

For each resource that promises to provide a service, any capacity payments should reflect the reasonably anticipated performance of that resource, where the "reasonably anticipated performance" should generally be based upon past performance. It is reasonable for resources to receive payments for providing ancillary services, even if the resources are not called upon to provide energy, provided that the payments reflect reasonable expectations of performance if called upon.

For load resources that provide energy through curtailments, the quantities of energy curtailed should be measured according to the best available estimates of loads in the absence of curtailment. Estimation of demand reductions according to historical baselines may be reasonable if the historical baselines reasonably reflect current loads in the absence of curtailment. Estimation of demand reductions according to loads just before and after a curtailment period may be reasonable under certain circumstances.

By developing an objective measure of performance, ERCOT can address a plethora of questions that have concerned ERCOT participants, such as the following:

- *How should interruptible loads of different sizes and load patterns be treated?* They should be paid for the services that they actually provide to the power system, which will depend upon their load characteristics and their abilities to respond to price or curtailment instructions.
- *Should temperature-sensitive loads be treated differently than other loads?* Given an appropriate measure of performance, all loads acting as resources should be treated identically.
- *Should loads that are around-the-clock operations be treated differently than other loads?* Given an appropriate measure of performance, all loads acting as resources should be treated identically.
- *Should loads bid "conservatively," offering a smaller quantity of service than they actually expect to provide?* This question is equally applicable to generators, because all resources face uncertainty in the quantity of service that they will actually provide. The measure of performance should reflect the fact that a resource that is more valuable if it

meets its promises than if it does not do so, that under-performance has a cost that may exceed the price paid for the promised level of service, and that over-performance has a value that may be less than the price paid for the promised level of service. Given appropriate measures of performance and prices that reflect the market's marginal cost of supply, each resource will find it profitable to manage its performance risk in a way that is consistent with offering the greatest possible net value to Texas power consumers.

- *Should payments to resources be capped so that they are not paid for more than 120% of the quantities reflected in their offers?* No. If over-performance has value, resources should be paid for that value. If the value of over-performance is lower than that of promised performance, the price paid for over-performance should be lower than that paid for promised performance.
- *Should the ancillary service markets for loads acting as resources be separate from those for generators?* No. This will make markets thinner, more volatile, and more subject to manipulation. The solution is to measure the performance of all resources on a consistent basis.
- *Should performance thresholds be established at the load level, the QSE level, or the ERCOT system level?* System reliability requires that performance thresholds be established at the system level. But if performance measures are blind to aggregation, the measured performance of all QSEs will sum to the performance of the system, and the measured performance of each QSE will equal the sum of the performances of the resources scheduled by the QSE.
- *To what degree can or should a QSE substitute generation and load resources for each other?* Given an appropriate measure of performance, such substitution should be permitted without limitation.

We therefore recommend that Texas develop methods for measuring the effective quantities of services provided by resources. These methods should follow the principles described above.

7.3. Long-Term Recommendations

In the long term, Texas needs an efficient pricing system that can serve as the basis for valuing and financing demand response programs. It will take time and effort to implement such a system. In particular, once Texas decides to do so, at least two or three years will be required for development of the software and hardware that will be needed to run such a system.

Our long-term recommendations can be divided into two sets. The first set of recommendations – concerning the introduction of centralized day-ahead and real-time markets, three-part bidding, and locational pricing – directly address the problem of efficient pricing. The second set of recommendations – concerning the application of market-clearing pricing, the levying of penalties, and the allocation of demand response program overhead costs – presume the existence of efficient pricing. Consequently, although it would be feasible to implement the second set before the first set, it would not be efficient to do so.

7.3.1. Develop Transparent Day-Ahead Electricity Markets

Except for Texas, all U.S. ISOs have (or soon will have) day-ahead markets in which participants can commit themselves to market-based hourly prices for producing power or conserving power. Because most loads can respond to price only with sufficient lead time, such day-ahead markets are an important tool for encouraging demand response. For many large industrial consumers of electricity, day-ahead notice of hourly prices is sufficient to allow rescheduling of their production shifts and to thereby allow response to power market conditions. For most of these same industrial consumers, same-day notice is not sufficient.

Texas would benefit from liquid day-ahead markets that would allow market participants to commit themselves to produce or consume power at known hourly prices. Such prices should vary by zone or location.

Because flows across potentially constrained interfaces depend upon the simultaneous actions of all market participants and because least-cost dispatch and efficient prices depend upon these flows, it is essential that the day-ahead market be operated by a single entity that has access to all available information about all transmission and other system constraints (e.g., voltages) and all transactions that affect flows through potentially constrained facilities. We understand that Texas seems to have a strong preference for allowing private firms to make its markets; and we do not assert that a private firm cannot operate a day-ahead market. We must note, however, that the system operator has the network data and the computer models needed to determine the simultaneous feasibility of proposed transactions and that any private firm making a day-ahead market would need to duplicate these data and models. We also note that the system operator's reliability obligation gives it a strong interest in assuring adequate unit commitment and its continual access to the necessary system information. Thus, as a matter of cost and duty, it would make sense to seriously consider having ERCOT serve as the day-ahead market maker, which is a role served by all of the other U.S. ISOs.

7.3.2. Develop Transparent Same-Day Electricity Markets

ERCOT's present system of bilateral trades, its balanced load requirement in the real-time market, and its transmission congestion makes it difficult for market participants to quickly trade the services that can be provided by load response. Such trades would be greatly facilitated by the creation of centralized same-day markets (such as hour-ahead and real-time markets) that would facilitate rapid trades among numerous participants, allowing each participant to find the best possible opportunities for buying and selling power. With prices that vary by zone or location, such markets would provide the timely information on available resources that is needed to allow QSEs to maintain their balanced schedules at least cost while respecting transmission constraints. This price discovery can be essential for allowing demand response to rapidly changing market conditions.

Because of the necessity of respecting transmission constraints, it is essential that same-day markets be operated by a single firm that has access to information on all same-day transactions so that it can determine the simultaneous feasibility of all proposed transactions and so that it can calculate the zonal or locational prices that are consistent with the day-ahead dispatch. As with day-ahead markets, the system operator is particularly well positioned to operate same-day markets, as it has the necessary computer models and has the obligation to assure reliable service. Indeed, all of the other U.S. ISOs operate same-day markets, all of which allow (or will allow) unbalanced schedules. For the same reasons applicable to day-ahead markets, Texas should seriously consider having ERCOT serve as the same-day market maker.

7.3.3. Consider Adopting Three-Part Bidding

Like generators, loads sometimes incur fixed costs when they provide electricity services. For example, some loads may incur high costs for even a brief curtailment but only moderate costs as the curtailment continues. Such loads will be reluctant to accept curtailments unless they are somehow assured that they will be compensated for the costs that depend upon the occurrence of a curtailment in addition to those costs that depend upon the duration of the curtailment.

To assure that resources recover both their fixed and variable costs of providing electricity services, the Northeastern ISOs allow resources to make three-part bids. For generators, these bids indicate the prices at which each generator is willing to start up and shut down, to operate at minimum load, and to provide energy. For loads, these bids can similarly indicate the fixed and variable costs of curtailments. In either case, a resource that would require a minimum "run time" under one-part (energy only) pricing will be willing to provide services for whatever period the system requires under three-part bidding. Furthermore, while one-part bidding requires the resource to guess their run times, to bid above incremental energy cost, and to risk under-recovery of costs, three-part bidding encourages bids that approximate incremental cost, without exposing resources to under-recovery risk. This has the benefit of encouraging bids from loads and generators who might otherwise stay out of the market; and it has the further benefit of making it easier to detect the exercise of market power by suppliers, as price-taking generators will bid near incremental cost.

7.3.4. Consider Adopting Efficient Locational Pricing of Electrical Energy

As noted above, transmission constraints and losses cause the value of energy services to vary by location. In order to capture the benefits of demand response in relieving transmission constraints, a mechanism is needed to communicate these locational values to consumers, or at least to the REPs who might offer demand responsive rates. Such communication is necessary both for inducing short-term response to power system conditions and longer-term investment in demand response.

If the locational value of power services were communicated to consumers through efficient locational prices, these prices would equal the market value of electricity at each system location. These prices are the *only* prices that are consistent with efficient system dispatch; and they are the *only* prices that induce self-interested loads to consume efficient quantities of power and profit-maximizing generators to produce efficient quantities of power.

A power system that does not have efficient locational prices must instead have *in*efficient locational prices. Zonal pricing is one such inefficient pricing scheme. Zonal pricing offers a single energy price to all market participants in a zone. At their best, zonal prices are averages of the efficient locational prices within the zone, so zonal prices are inevitably inefficiently high in some locations and inefficiently low in other locations. Zonal pricing thus makes the price of electricity too low in precisely those locations where demand response and increased generation output are most valuable. By failing to recognize the fact that demand response and new generation have higher values in some locations than in others, zonal pricing induces market participants to behave in ways that increase congestion costs.

Zonal pricing has been tried and abandoned by California, New England, and PJM. It was tried because market participants expected that it would simplify the trading process. It failed because, in spite of its apparent simplicity, zonal pricing in fact complicates system control and especially congestion management:

- It leads market participants to invest in demand response and generation in the wrong locations, exacerbating congestion.
- It makes it necessary for system operators to make special payments to consumers and generators who can relieve supply shortages in load pockets, and can lead to increasingly complex administrative programs as loads pockets change from hour to hour and year to year.
- It makes system dispatch unnecessarily costly both within zones and between zones.
- It leads some traders to intentionally create intra-zonal congestion so that they can be paid to relieve such congestion.

In short, zonal pricing dramatically increases the complexity of congestion management, requires the creation and continual revision of administrative programs to manage congestion, increases system costs, and undermines (rather than enhances) market processes. With zonal pricing, Texas must create a variety of programs to counteract the bad pricing signals inherent in zonal pricing; and Texas is in fact doing so because it recognizes that it has load pocket problems.[51] Unfortunately, these programs are and will be primarily administratively determined, rather than market-driven. With efficient locational pricing, such programs would be unnecessary.

7.3.5. Apply Market-Clearing Prices to All Resources' Effective Quantities

For each service in each time interval, all load and generation resources should receive the same price per unit of the *effective quantity* provided. These prices should be differentiated by zone or location.

7.3.6. Set Penalties for Non-Performance Equal to the Costs of Non-Performance

Penalties for non-performance are unnecessary and counterproductive, except to the extent that they reflect the power system's costs of non-performance. Penalties are unnecessary because it is sufficient for loads and generators to receive market prices for the services that they provide and to pay market prices for the services that they use. It is counterproductive because there are many situations in which it would be better for the Texas economy for some market participants,

[51] At least three documents refer to the problem and programs mentioned in the text. ERCOT Task Force on Demand Side Resources and Demand Responsiveness, *Issues Pertaining to the Balancing Up Load Market*, undated, p. 6, asks, "Should there be special BUL programs in certain critical transmission-constrained areas (e.g., the DFW area)?" Task Force on Demand Side Resources and Demand Responsiveness, *Preliminary List of Issues that Might Be Explored*, August 2, 2001, p. 2, and J. Zarnakau, *The Role of Load Management and Demand-Side Resources in ERCOT*, July 2001, slide 29, both include the following identical statement of an outstanding issue: "Determine whether it is possible to establish a minimum annual payment for BUL resources in certain critical areas (Dallas-Fort Worth, for example). This could involve treating BULs similarly to reliability must-run (RMR) resources in these critical regions, with guaranteed minimum payments."

with relatively high-cost resources, to "lean" on spot markets so that they obtain services from market participants with relatively low-cost resources.

In general, the relevant "costs" are the power system's cost of providing the energy and ancillary services needed to compensate for the non-performance. These costs can usually be measured according to the market-clearing prices of the needed energy and ancillary services. In extreme cases, the costs may also include the harder-to-quantify costs of the reliability risks created by the non-performance.

The key to providing efficient performance incentives to market participants is to accurately measure the services provided by and used by each market participant. Texas' electricity consumers should be indifferent to resources providing more or less of their promised services if the payments for these services accurately reflect both the marginal cost of alternative resources and the quantities of services actually provided. Sometimes the costs of resolving the reliability and frequency control problems caused by non-performance can be extremely high, in which case the non-performers should pay correspondingly high prices for non-performance.

The capacity payments that resources receive for promising to provide ancillary services should reflect their performance in actually providing power upon the ISO's request. The "penalty for non-performance" should generally be an adjustment in each resource's quantity promises to reflect their record of actual performance, and to reflect the costs to ERCOT of their non-performance.

Under the present Protocols, loads acting as resources have an advantage over generators because generators are penalized for uninstructed deviations whereas loads are not. In general, however, loads and generators should be treated equally, both paying for the costs of their uninstructed deviations but not further penalized.

The risks of non-performance should be borne by the load or generator resource that fails to perform rather than by ERCOT and ERCOT's other customers. Some market participants have argued that load uncertainties are often beyond the control of loads acting as resources and that these loads should have "safe harbors" that exempt them from penalties for non-performance. Certainly, ERCOT should do its best to provide loads with advance information concerning how they will be compensated for performance – information such as on the prices that they will receive and on the baseline loads upon which performance will be measured. But because each resource's risks of non-performance are best managed by that resource, the costs and benefits of managing those risks should be borne by that resource alone.

ATTACHMENT 1.
ERCOT DOCUMENTS REVIEWED

Anonymous:

- *An Overview of the Restructured Texas Electricity Market*, undated.

Demand-Side Working Group of the Wholesale Market Subcommittee:

- *Meeting Notes from the March 18, 2002 Meeting: Interpretation of Current Protocol Requirements for Loads Acting as Resources Providing Responsive Reserves*.
- *What Loads Should Know About the ERCOT Protocols Before They Read Them*, Draft 1.0, June 3, 2002.

ERCOT Protocols:

- Section 4, *Scheduling*, May 1, 2002.
- Section 5, *Dispatch*, April 1, 2002.
- Section 6, *Ancillary Services*, May 6, 2002.

Jerry Golden Energy Services, presentations:

- *Making the Balancing Up Load (and other small) Markets Work*, August 2001.

Task Force (Working Group) on Demand Side Resources and Demand Responsiveness:

- *Issues Pertaining to the Balancing Up Load Market*, undated.
- *Issues Pertaining to Interruptible Loads and Loads Acting as Resources*, undated.
- *Meeting Notes: August 29th, 2001*, August 30, 2001.
- *Meeting Notes: December 13th, 2001*, December 18, 2001.
- *Meeting Notes: February 25, 2002*, March 3, 2002.
- *Meeting Notes: May 6, 2002*, May 30, 2002.
- *Meeting Notes: October 19th, 2001*, October 24, 2001.
- *Meeting Notes: November 8th, 2001*, November 8, 2001.
- *Notes from August 14, 2001 Kick-Off Meeting*.
- *Preliminary List of Issues that Might Be Explored*, August 2, 2001.
- *Status Report*, January 2002.

Jay Zarkinau, papers:

- *Promoting Greater Demand-Side Resource Participation in ERCOT's Ancillary Services Markets Without Compromising Reliability*, April 19, 2002.

Jay Zarkinau, presentations:

- *Measuring the Quantity of the Resource Provided by a Load Acting as a Resource in the Responsive Reserve Market*, March 2002.
- *Selling an Interruptible Load Resource or Demand response in ERCOT's Restructured Market*, May 2002.
- *The Role of Load Management and Demand-Side Resources in ERCOT*, July 2001.
- *Transitioning ERCOT's Demand-Side Resources into the New Market: A Scorecard*, December 2001.